Semantic Web構築に向けたメタデータ自動作成に関する研究

Semantic Web構築に向けたメタデータ自動作成に関する研究

岡野拓生（B4）†，　鈴木育男‡，　久保洋†
†室蘭工業大学　情報工学科　　‡室蘭工業大学　SVBL

1. はじめに

ネットワーク空間での情報分散化を重視したハイパーテキストシステムの一つであるWebは，記述の容易さやコンピュータの普及，インターネット人口の増加などから爆発的に普及し，誰もがアクセスできる巨大な情報共有空間となっている．毎日の天気や最新のニュースなど，Webから提供される情報は，我々の生活にはなくてはならない社会的基盤の一つとなってきていると考えられる．

しかし，Web上に構築される情報空間が巨大になるにつれて，ハイパーリンクを辿る事によって情報を獲得するだけでは，目的とする情報を獲得するまでに多くの労力を必要とし，効率が悪い．そのため，この情報検索の労力を減少させるために，これまでにYahooやGoogleといった情報検索サービスが提供されてきた．このような情報検索サービスでは，検索のための入力文字の選択が検索効率に大きな影響を与えている．つまり，検索サービスを使用するユーザが，適切な検索文字を入力しないと目的の情報が入手できないことになる．これは，現在使用されているWeb上の情報記述言語であるHTMLが，文字サイズや段組などの装飾情報しかもっていないことが原因であると考えられる．Web上の情報に，その意味や概念等の情報を付加できれば，その概念情報を基によりよい情報提供が可能となる．

このような問題を解決するための一つの方法として，「セマンティックWeb」が提唱されている．このセマンティックWebでは，これまでのHTML文書を処理するばかりではなく，その情報の意味内容を表現するメタデータを付加し，このメタデータに対して様々な情報処理を行うことでWeb上の情報を効果的に利用しようという試みである．またメタデータ自体は，HTML文書以外の情報にも関連づけることが可能であるため，画像・動画・音楽といった文字以外の情報の活用も可能になる．

セマンティックWebでは，メタデータを基本としているので，メタデータ情報が普及しなければセマンティックWebの実現はできない．セマンティックWebの実現に向けたもっとも根本的な問題として，「誰がメタデータを書くのか」ということがあげられる．これは，これから新しく発信される情報に対してメタデータを付加するということというよりは，むしろ，現在Web上に存在している膨大な情報にメタデータを付加して再構築するかという問題のほうが大きいように感じる．

そこで，本研究ではセマンティックWebの実現に向けて，メタデータを容易に付加できるシステム環境を構築することを目的とする．具体的には，現存するHTML文書からメタデータを必要とする部分を抜き出し，メタデータ用のタグ情報を自動的に追加するものとする．

2. セマンティックWeb技術と関連研究

セマンティックWebでは，前述したようにWebページ（HTML文書）にそのページの内容や関連した情報について，機械的な処理が可能なメタデータを付加することにより，情報の質をコントロールすることを目指している．さらに，付加したメタデータの意味や相互関係をオントロジーによって関連づけることで，メタデータ間の概念関係を表現可能となり，それによって表層語の違い（日本語と英語の違いなど）を補うことができる．

セマンティックWebにおいて，メタデータの記述言語には，RDF(Resource Description Framework)が標準化されている．RDFの意味モデルは，主語（subject），述語（predicate），目的語（object）の3つの要素により構成されており，この主語と目的語を述語で結合する有向グラフとして表現される．例として，『お好み焼き』の『材料』が『キャベツ』であることをRDF意味グラフでは，図1の有向グラフで表現される．

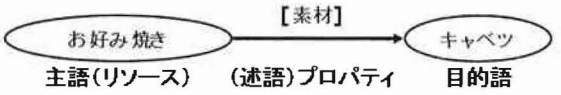

図1：RDFモデル

このように，RDFを利用すれば，Web上に意味ネットワークを構築できるのであるが，RDFの作成は容易な作業ではない．現在，メタデータ自動作成機能としては，南野ら[1]の「何でもRSS」，「MyRSS」[2]などにおいて，RDFの機能を絞ったRSSの記述に関して研究が行われている．

3. 提案手法

本研究では，HTMLからRDF記述を自動的に行えるシステムの構築を試みる．構築システムの処理作業は以下のようになる．

1. HTML文書を形態素解析によりメタデータを付加する要素を抽出
2. 抽出した要素に専用のタグを付加
3. タグを基にRDFテンプレートに記述

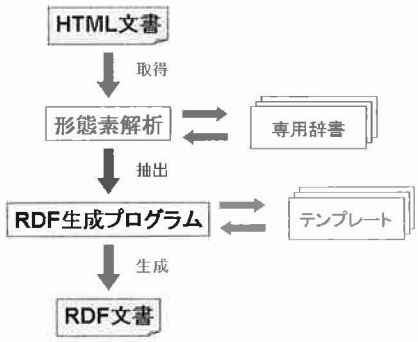

図２：提案手法

3.1. 形態素解析による要素抽出（作業１，２）

形態素解析とは，入力された文字列を単語ごとに分解し品詞などの文法情報や意味情報など付加する処理である．本研究では，形態素解析のソフトとして『茶筌』を使い，単語ごとに分解し，自作した専用辞書を利用して，意味情報を付加する．

この際，意味情報を表現するために＜タグ＞を利用する．HTML や XML と同様に，タグで囲むことにより，HTML パーサでの処理を可能とする．

3.2. RDF テンプレートの利用（作業３）

形態素解析での処理作業を通じて，必要な単語要素をHTML と同様なタグで囲んだことによって，HTML パーサを使用した解析処理が可能となった．パーサの役割としては，タグごとに分解された文章のタグが，形態素解析によって付加されたタグであるかのチェックを行い，タグで挟まれている要素をメタデータの目的語の候補として選択することである．また，囲まれているタグの種類によって，その要素がどの述語の目的語であるかも定義可能となる．

本研究では，パーサによって切り出したタグ情報を，基本となるRDFテンプレートに挿入することで新たにRDF ファイルを作成する．

4. 検証実験

4.1. 実験準備

構築した RDF 自動作成システムについて，検証実験を行った．現在使われている RDF には，Blog 上での RSS や DublinCore による書籍情報の管理などがあるが，本研究では，料理情報の管理について RDF の自動作成を行う．

まず，料理情報に関する形態素解析用の辞書を作成した．形態素解析により付加するタグの意味情報は以下の4つを用意した．

- <ryouri>　～ 料理名
- <sozai>　～ 料理に使用される素材名
- <fumi>　～ 料理の風味の種類
- <tani>　～ 材料単位

これらのタグと RDF テンプレートを利用して，料理管理用の RDF ファイルを作成した．

4.2. 実験結果

実証実験に使用した Web ページを図３に示す．

図３：処理前の Web ページ

これから，形態素解析，パーサ，テンプレートへの埋め込みの各作業を通じて作成された RDF ファイルの一部を以下に示す．

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:eg="C:/sample/test.html">
    <rdf:Description rdf:abut="C:/sample/test.html">
      <eg:ryouri xml:lang="ja">お好み焼き</eg:ryouri>
      <eg:fumi xml:lang="ja">和風</eg:fumi>
       <eg:image><imgsrc="http://image.infoseek.co.jp/
            recipe/I/199611326.jpg"width="220"height="169"border=
            "0"alt="広島風
      </eg:image>
      <eg:cal>723kcal</eg:cal>
      <eg:sozai rdf:parseType="Collection">
        ・・・・(中略)・・・・
        <rdf:Description xml:lang="ja" rdf:abut="#キャベツ"/>
        <eg:tani xml:lang="ja">550g</eg:tani>
        <rdf:Description xml:lang="ja" rdf:abut="#もやし"/>
        <eg:tani xml:lang="ja">160g</eg:tani>
        ・・・・(中略)・・・・
      </eg:sozai>
    </rdf:Description>
</rdf:RDF>
```

5．おわりに

本研究では，これまでの情報処理の流れに大きな変革をもたらすであろう，セマンティック Web の容易な導入のために，RDF の自動作成ということについて議論した．そして，現在ある Web 上の情報（HTML）から，形態素解析や XML パーサなど既存の技術を利用して，RDF を自動作成可能なことを検証実験により示した．

参考文献

[1] 南野朋之，鈴木泰裕，藤木稔明，奥村学，“blog の自動収集と監視，” 人工知能学会誌，Vol.19，No.6，pp.511-520，2004.

[2] “MyRSS.jp，” http://myrss.jp/ .